DOCUMENT RESUME

ED 063 338 TM 001 356

AUTHOR Koehler, Roger A.
TITLE Coombs' Type Response Procedures.
PUB DATE Apr 72
NOTE 13p.; Paper presented at the annual meeting of the American Educational Research Association (Chicago, Illinois, April 1972)

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS Comparative Analysis; *Guessing (Tests); *Objective Tests; *Response Style (Tests); *Test Construction; *Testing; Test Reliability; Test Results; Test Validity
IDENTIFIERS Coombs (C H); Dressel (P L); Schmid (P)

ABSTRACT

This paper provides substantial evidence in favor of the continued use of conventional objective testing procedures in lieu of either the Coombs' cross-out technique or the Dressel and Schmid free-choice response procedure. From the studies presented in this paper, the tendency is for the cross-out and the free choice methods to yield a decrement rather than an improvement in the quality of measurement characteristics when compared to conventional testing procedures. (Author/DB)

ED 063338

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY

Coombs' Type Response Procedures

Roger A. Koehler
Ohio University

TM 001 356

The objective examination has come under heavy criticism for many years. One reason for this criticism is the belief that objective tests are not accurate measures of the traits they are designed to test. For example, scores on these tests may be particially determined by an examinee's guessing behavior when he is not sure of the correct answer to a particular item. Various methods have been proposed over the past fifty years to remove the effects of guessing from test scores, but at the present time there does not appear to be an ideal solution to the problem.

Stokes (1966) presented an excellent illustration of the contaminating effects of guessing when he described a hypothetical student on "test day". He wrote:

> As expected of a typical student, Mary arrives for class reasonably well prepared. Armed with a blend of knowledge and courage, she takes the test paper in her somewhat unsteady hand and reads the first of a long list of multiple-choice questions. Years of exposure to such tests lead her to seek the best or most appropriate answer. There is a pause of deliberation and then the fateful mark. What are the alternatives?
>
> - If Mary knows the answer, there is little doubt that she will mark the correct response and move on to the next question with dispatch.
> - If Mary is uncertain, the process is quite different. After the first frozen instant, she leans with all her might against the inner logic of the question, hoping to find the answer. Each alternative is read and reread, reversed and held in mental mirror image to find its subtle truth. Each possibility is weighed against the others and against the consequence of missing the item. Tension comes uninvited. Slowly, reluctantly, choices are put aside in a significant emotional drama, as she hopefully but fearfully funnels them to an acceptable response. Then, one mark and only one: a guess - perhaps an educated and considered one - but still a guess. (p. 271)

It is clear that little if any useful information as to her knowledge of the subject matter can be obtained from the guessed response. If she guesses correctly on a particular item, one point is added to her

test score; a point that is not an indication of what Mary knows about the subject matter being tested (validity suffers). On the other hand, if Mary guesses incorrectly, she is given no credit for her response when in fact she may have been able to eliminate some of the alternatives. In other words, she may not be completely lacking in knowledge, but may have partial information.

The last statement suggests a possible solution to the problem of guessing on objective tests; i.e., the assessment of varying degrees of partial knowledge. In light of this, the examinee would no longer be forced to make a single response (select the correct alternative), to an item, but could express various levels of partial information when he was not sure of the correct answer.

One possible technique for measuring partial knowledge was introduced by Coombs (1953). The test directions for the Coombs technique instruct examinees to cross out all the alternatives they believe are incorrect and to avoid guessing from the remaining alternatives. If, indeed, an examinee is certain of the correct answer, he would cross out all the incorrect alternatives and would receive one point for each alternative crossed out. On the other hand, if the person taking the test was not certain of the correct answer, he could still receive partial credit for crossing out one or more alternatives he recognizes as incorrect (one point for each alternative). Under Coombs' directions, if an examinee mistakenly crosses out the correct alternative, he is assessed a penalty of 1 - k points where k is the number of alternatives per item. Therefore, it would be unwise for an examinee to guess when he is not certain an alternative is incorrect.

As an example of the above procedure, consider a test with five alternative items. If a person is sure that alternative -2- of a particular item is correct, he would cross out alternatives -1-, -3-, -4-,

and -5- and would receive four points for that item. Suppose this person does not know the answer to another item, but recognizes that alternatives -3- and -5- are absurd. He would then cross out those two alternatives and would receive two points of partial credit. Now, if this examinee is totally uncertain as to which alternative is correct for a particular item but decides to guess that alternative -3- is incorrect, the fact that alternative -3- is the right answer would result in a score of -4 for that item. If for some reason a person taking the above test decided to cross out all five alternatives to an item, he would receive the same score (4 + 1 - 5 = 0) as if he had left the item blank.

Coombs, Milholland, and Womer (1956) performed an experiment to determine the possible merits of the Coombs' testing procedure. They administered three different tests to 855 high school students, under both conventional (select the one best answer) and Coombs' directions. An experimental design was developed such that no student received the same test under more than one type of directions. A third testing method was employed to determine whether or not partial information enters into the responses on multiple-choice tests. Under this third method, students were first asked to rank three out of the four alternatives as to their attractiveness as distractors. Then, they were asked to respond to the items under Coombs' directions (i.e., circle those alternatives you are certain are distractors). Assuming that the alternative ranked third would be the students' second choice for the correct answer to an item, the proportion of the alternatives that were so ranked and were correct answers was used as a measure of partial information. This proportion was found to be greater than would have been expected by chance and for this reason, the authors concluded that partial information does, indeed, enter into multiple-choice test scores.

The second phase of the Coombs' et al. study compared the reliability and validity of Coombs' testing procedure with that of the conventional method. Using Kuder-Richardson-20 reliability estimates, it was found that the application of Coombs' directions yielded an increase in reliability equivalent to a 20 per cent increase in test length. A total of sixteen criterion variables were available for calculating validity coefficients. These variables consisted of various sub-scores of the Differential Aptitude Test, the Stanford-Binet IQ Test, the California Intelligence Test, the Detroit Aptitude Test and the Mac Quarrie Mechanical Aptitude Test. Although none of the validity coefficients were reported in the study, the authors concluded that no significant change occurred when Coombs' directions were employed.

Finally, Coombs et al. suggested that the Coombs testing method may introduce a new dimension into the overall variance of test scores. While, supposedly, the guessing component has been eliminated by this new testing procedure, the authors suggest that students may differ in the threshold of assurance at which they would respond to test items.

In a recent study (Collet, 1971), reliability and validity comparisons between the Coombs' response procedure and corrected for guessing conventional response scores were examined. Collet found no significant differences in reliability between the two methods; however, a comparison of criterion related validity coeffecients yielded superior results for the Coombs' method over corrected for guessing scores. Numerous studies (e.g., Jackson,1955;Little,1966; Mead and Smith,1957) have shown that the standard correction for guessing over-corrects or under-corrects with the presence of partial information and misinformation. Other studies

(Sheriffs and Boomer, 1954; Votaw, 1936) suggest that the correction formula used with do-not-guess directions introduces a personality dimension into test scores. To quote Slakter (1967):

> "The implication for the individual taking a test under conventional directions and the usual penalty for guessing is clear: Answer all questions! The implication for the test constructor is also clear: If you include a penalty for incorrect responses, the test scores of the examinees will reflect their RTOOE [risk-taking] strategies as well as their aptitudes or achievements!" (p 43)

Still further studies (e.g., Michael, Stewart, Douglas, and Rainwater,1963; Patterson and Langlie, 1925; Ruch and DeGraff, 1926) provide inconsistent results concerning the improvement of test characteristics (reliability and validity) through the use of the standard correction for guessing. In light of the above studies, one must wonder whether the comparison of Coombs' procedure with the conventional method corrected for guessing is worth the effort.

A test taking technique similar to that of Coombs (1953) was proposed by Willey (1960). In order to obtain information on as examinee's ability to discriminate among the incorrect alternatives, Willey instructed the examinees to indicate the two alternatives out of five that they believe are least likely to be correct. The examinees were also told to select the one best answer to each item. Item scores were obtained in the following manner:

Keyed alternative selected as best answer = 1

Keyed alternative not marked at all = -1

Keyed alternative selected as being wrong = -3

The rationale behind this testing procedure is essentially the same as was reported for the Coombs' method. However, rather than giving positive

points for partial information, Willey assessed fewer penalty points.

Bernhardson (1966, 1967) indicated that no improvement in test characteristics was evidenced for Willey's procedure over conventional rights-only objective testing. Bernhardson (1966) found that the chance score on a 20 - item multiple-choice test was approximately double that of conventional testing when Willey's procedure was utilized. In addition, Bernhardson (1967) correlated scores obtained through the Willey procedure and through the conventional method with end of term academic average. The results showed no significant differences in the correlations for the two response methods.

Another somewhat different variation of the Coombs technique was proposed by Dressel and Schmid (1953). The latter authors asked examinees to select as many alternatives as they considered necessary to make sure they had marked the correct alternative. This response procedure was referred to as the free-choice method, and the item score was determined by the number of choices the examinee made. For an item with k alternatives, the item score would be k-n if the correct alternative was among the n selections made by the examinee. On the other hand, an item score of -n would result if the correct alternative was not among the n selections.

Dressel and Schmid (1953) investigated the effect on reliability and validity of their free-choice response technique as compared to that of the conventional choice procedure. They first administered a typical multiple-choice test to ninety college students. Immediately following the first administration, the students were given the same test under free-choice directions. An analysis of the scores on the two administrations indicated that the free-choice test was inferior to the conventional choice test with respect to both reliability and validity.

Two extensive studies (Moore, 1956; and Archer, 1962) were performed to compare the reliability and validity of (1) Coombs' cross-out (CO) method, (2) Dressel and Schmid's free-choice (FC) technique, and (3) conventional choice testing (CT). The remainder of this paper will present a discussion of the results of these two studies.

Reliability

Moore (1956) administered a vocabulary test to three different samples, each containing over 350 ninth grade students. Each sample was given the test under a different response procedure (CO, FC, or CT). The amount of time allowed for completing the test was 10 minutes, regardless of the response procedure employed. Referring to Table 1, it can be seen that split-half reliabilities tended to be higher in the CT group than in either the CO or FC group. The only statistically significant difference in reliability occured between CO scores and CT scores when an estimate of reliability per unit of testing time was colculated (i.e., reliability based on a time limit of 10 minutes).

TABLE 1

Reliabilities from the Moore and Archer Studies

	CONVENTIONAL	CROSS-OUT	FREE-CHOICE
Moore			
Reliability/10 min.	.87	.75	.81
Equivalent Length	.93	.86	.89
Archer			
Parallel Forms	.85	.84	.82
KR #20			
Form A	.84	.86	.80
Form B	.88	.87	.82

Moore also calculated split-half reliabilities for high, average, and low ability groups separately. The results for these ability groups were essentially the same as for the total group and, hence, are not reported in this paper.

Archer (1962) suggested that the negative findings of the Moore (1956) study might be attributable to a lack of training for students responding under the CO or FC methods. He, therefore, spent over one hour teaching students the intricacies of responding to test items under the CO or FC directions. Following this training period, students were administered two forms of a 60 - item social studies test under the response procedure they had learned. Approximately one-third of a total of over 500 sixth grade students were given the test under the CO method, one-third under the FC method, and one-third under the CT procedure. Table 1 indicates that no significant differences in parallel forms reliability coefficients resulted among the three response techniques.

Archer suggested that perhaps the difficulty of the items is related to the success of the CO and FC response techniques. He, therefore, considered the reliability of the three response procedures for "easy" items, "average" items and "difficult" items. Since the results for the various difficulty levels were similar to those of the total test, these separate reliabilities are not reported here.

Validity

Moore (1956) used scores on the Iowa Tests of Basic Skills - Vocabulary (ITBS-V) and total composite scores on the Iowa Tests of Educational Development (ITED) as outside criteria for validity considerations. Table 2 presents the correlations between Moore's vacabulary test and

each of the two outside criterion for the three different response procedures (CO, FC, and CT). No significant differences in validity coefficients resulted among the three response systems.

TABLE 2

Validity Coefficients from the Moore and Archer Studies

	CONVENTIONAL	CROSS-OUT	FREE-CHOICE
Moore			
ITBS-V	.84	.90	.86
ITED	.81	.72	.84
Archer			
Teacher Ranks			
Form A	.63	.50	.59
Form B	.66	.55	.58
ITBS			
Form A	.78	.71	.61 *
Form B	.76	.73	.56 **

* Conventional -vs- Free-Choice is significant at .05 level
** Conventional -vs- Free-Choice is significant at .01 level

While the correlations between vacabulary scores and the ITBS-V tended to be slightly higher in the CO and FC groups than in the CT group, the corresponding correlations with the ITED did not yield such tendencies.

The outside criteria in Archer's (1962) study were composite grade equivalent scores on the ITBS and teacher rankings of social studies ability. An inspection of Table 2 indicates a tendency for CT scores to yield slightly higher correlations with teacher rankings than does either CO or FC responses; however, not significantly so. In addition, FC responses tended to be inferior (not significantly) to CO responses when correlated with teacher rankings.

Using the ITBS as a criterion, one can see from Table 2 that FC scores

yielded significantly lower validity correlations than CT responses. However, no significant differences in social studies vs. ITBS correlations occurred between the CO and CT groups.

As was true for reliability comparisons, the difficulty level of the test items did not appear to affect the validity considerations. Therefore, the results for various difficulty levels are not presented here.

In summary, this paper provides substantial evidence in favor of the continued use of conventional objective testing procedures in lieu of either the Coombs' cross-out technique or the Dressel and Schmid free-choice response procedure. To this writer's knowledge, there is no data to suggest that the use of these latter two non-conventional response methods will improve the reliability or validity of the objective examination. If any trend were to be extracted from the studies presented in this paper, it would be the tendency for the CO and FC methods to yield a decrement rather than an improvement in the quality of measurement characteristics when compared to conventional testing procedures. The only evidence in the literature in support of the CO technique compared the CO technique with the CT method where CT scores were corrected for guessing. Since the literature tends to discredit the use of the standard correction for guessing, a comparison such as the above does not appear to be worthwhile.

References

Archer, N. S. A comparison of conventional and two modified procedures for responding to multiple-choice test items with respect to test reliability, validity, and item characteristics. Unpublished doctoral dissertation, Syracuse University, 1962.

Bernhardson, C. S. Determination of the chance score on the three-decision multiple-choice test. Psychological Reports, 1966, 19, 559 - 562.

Bernhardson, C. S. Comparison of the three-decision and conventional multiple-choice tests. Psychological Reports, 1967, 20, 695 - 698.

Coombs, C. H. On the use of objective examinations. Educational and Psychological Measurement, 1953, 13, 308 - 310.

Coombs, C. H., Milholland, J. E. & Womer, F. B. The assessment of partial knowledge. Educational and Psychological Measurement, 1956, 16, 13 - 37.

Collet, L. S. Elimination scoring: an empirical evaluation. Journal of Educational Measurement, 1971, 8, 209 - 214.

Dressel, P. L., & Schmid, P. Some modifications of the multiple-choice examination. Educational and Psychological Measurement, 1953, 13, 574 - 595.

Jackson, R. A. Guessing and test performance. Educational and Psychological Measurement, 1955, 15, 74 - 79.

Little, E. B. Overcorrection and undercorrection in multiple-choice test scoring. Journal of Experimental Education, 1966, 35, 44 - 47.

Mead, A. R. & Smith, B. M. Does the true-false scoring formula work? some data on an old subject. Journal of Educational Research, 1957, 51, 47 - 53.

Michael, W. B., Stewart, R., Douglass, B. & Rainwater, J. H. An experimental determination of the optimal scoring formula for a highly-speeded test under different instructions regarding scoring penalties. Educational and Psychological Measurement, 1963, 23, 83-99.

Moore, R. A. A comparison of selected modifications of a multiple-choice examination. Unpublished doctoral dissertation, State University of Iowa, 1956.

Paterson D. G. & Langlie, T. A. Empirical data on the scoring of true-false tests. Journal of Applied Psychology, 1925, 9, 339 - 348.

Ruch, G. M. & DeGraff, M. H. Corrections for chance and "guess" vs. "do not guess" instructions in multiple-response tests. Journal of Educational Psychology, 1926, 17, 368 - 375.

Sheriffs, A. E. & Boomer, D. S. Who is penalized by the penalty for guessing. Journal of Educational Psychology, 1954, 45, 81 - 90.

Slakter, M. J. The measurement and effect of risk taking on objective examinations. Final Report, Project No. 58428, Contract No. OE-6-10-239. U. S. Department of Health, Education, and Welfare, 1967.

Stokes, R. R. The split-response technique. Phi Delta Kappan, 1966, 47, 271 - 272.

Votaw, D. F. The effect of do-not-guess directions upon the validity of true-false and multiple-choice tests. Journal of Educational Psychology, 1936, 27, 698 - 703.

Willey, C. F. The three-decision multiple-choice test: a method of increasing the sensitivity of the multiple-choice item. Psychological Reports, 1960, 7, 475 - 477.